PRE-SYNOPSIS OF THE PROPOSED DOCTORAL RESEARCH WORK

ON

Development of Novel Deep Learning Methods for Deepfake Image Detection

For the Award of

Doctor of Philosophy in

Computer Science

Faculty of Science & Technology

Submitted By

Mrs. Sudha Pradeep Patil

M.Sc. (Comp. Sci), SET

Under the Supervision of

Prof. Dr. Sudhir B. Jagtap

Principal & Research Supervisor

Swami Vivekanand Mahavidyalaya, Udgir

December 2024

CONTENT

Sr. No.	Content Details	Page No
1	Introduction	1
2	Objective of the Study	3
3	Scope of the present Research	3
4	Literature Review	3
5	Statement of Research Problem	7
6	Hypothesis	7
7	Research Methodology	8
8	Significance of the Research	8
9	Data Collection Method	9
10	Data Analysis	9
11	Tentative Chapter Plan	10
	Chapter 1: Introduction	
	Chapter 2: Theoretical Background and Literature Review	
	Chapter 3: Problem Statement and Objective	
	Chapter 4: Research Methodology	
	Chapter 5: Data Interpretation and Analysis	
	Chapter 6: Findings, Suggestions and Conclusions	
12	References and Bibliography	11
13	Working plan	13

1. Introduction

With the rise of low-cost and powerful smart devices like mobile phones, desktop computers, and laptops, creating and editing multimedia content such as audio, photos, and videos has become easy and accessible. The widespread use of social media further enables the rapid sharing of creative content online, making it readily available to everyone.

The advancement of technology and Artificial Intelligence (AI) has made it possible to create realistic yet fake audio, images, and videos using deep learning techniques. Deepfakes are AI-generated fake media that mimic a person's likeness. The term "deepfake" is derived from the words "deep learning" and "fake," with deep neural networks forming the foundation of this technology [1]. The term was first popularized in late 2017 by a Reddit user [2], [3].

Deepfake content often involves swapping human faces, mimicking blinking, lip-syncing, and head movements to generate realistic images or videos of a target person [1], [4]. Applications like FaceApp (https://www.faceapp.com/), Deepfake Web (https://deepfakesweb.com), Zao (https://zaodownload.com/), and audio-focused apps such as Resemble (https://www.resemble.ai) facilitate users to perform face swaps, alter appearances, and even create fake pornographic videos [2] Some startups, such as Synthesia (https://www.synthesia.io/), use advanced tools to generate deepfakes, including entirely synthetic personas [4] The proliferation of deepfake synthetic media on the internet has made it widely accessible to the public.

Deepfake technology is positively used in various fields such as education, entertainment, brand personalization, games, digital communication, social media, healthcare, material sciences, and various business fields such as fashion and e-commerce. It can transform marketing and advertising in a significant way. Still, it has nefarious motives; it is used for spreading misinformation and generating pornography videos that hurt the public emotions. Fake videos have also led to political tension [5]and financial fraud. They have been used to degrade, intimidate, and harass not only celebrities, politicians, and CEOs, but also everyday citizens and pose reputational harm and create societal implications[[6], [7].

Figure 1. Example of deep fake video, the first row is the original frame, and second row contains generated fake frame.

With the increasing interest in deepfake technology, more and more research has been conducted. In the research on the detection of deepfakes using Deep Learning, the researchers' primary focus is to develop an algorithm that detects altered images or videos. Researchers are working on developing new datasets containing real and false pictures or videos for training and evaluation.

The most famous example of a deepfake is the video released in 2018 by Buzzfeed based on a synthetic Barack Obama and a genuine Jordan Peele.

Figure 3. Jordan Peele narrated Buzzfeed's Synthetic Obama video. [https://youtu.be/cQ54GDm1eL0] [5], [8]

2. Objective of the Research:

1. Identify the gap in the literature by a comprehensive survey of existing deepfake detection techniques, focusing on deep learning approaches.
2. Gather a dataset that includes both actual and fake images/ video in appropriate proportions.
3. Develop a robust deep fake detection model.
4. Compare and analyse the findings from proposed model with existing system Test and assess the model's performance with various parameters.

3. Scope of the present Research:

The goal of this research is to detect whether an input video is original or artificially made using deep learning approach.

4. Literature Review:

Deepfakes pose a danger to the privacy, integrity, and dependability of media material and consequently, it has attracted significant interest from researchers.

Yogesh Patel et al. discussed the generation and detection of deepfakes, highlighting advancements in machine learning and deep learning techniques, particularly using GANs.

Asad Malik[6], in a survey discusses the advancements and challenges in DeepFake detection, highlighting the use of deep learning techniques to identify manipulated multimedia content. It outlines various detection methods, datasets, and the implications of DeepFakes in media and cybersecurity while emphasizing the need for robust detection strategies to combat misuse.

Deng Pan [9] achieved impressive accuracy rates of 91-98% using the FaceForensics++ dataset using Xception and MobileNet architectures.

David Guera[2]presented a method for detecting deepfake videos using a temporal-aware pipeline that combines a Convolutional Neural Network (CNN) for frame-level feature extraction and a Recurrent Neural Network (RNN) for temporal

analysis. With the rise of accessible deepfake creation tools, the potential misuse of manipulated videos in various harmful scenarios is alarming. The proposed system is evaluated on a dataset of 600 videos, achieving a detection accuracy of 94% compared to random detectors. The research highlights the importance of understanding the manipulation process, which introduces temporal inconsistencies that the detection system can exploit. The findings suggest that even short video segments can be effectively analyzed for deepfake content, paving the way for improved digital media forensics.

Akash Chintha [10] presented advanced digital forensic techniques aimed at detecting audio spoofing and visual deepfakes, which are becoming more prevalent in manipulating public perception. The authors introduce a method that combines convolutional latent representations with bidirectional recurrent structures that are further enhanced by entropy-based cost functions. This innovative approach is designed to effectively capture both the spatial and temporal characteristics of deepfake content. The effectiveness of these methods is demonstrated through testing on the FaceForensics++ dataset, Celeb-DF video datasets, and ASV Spoof 2019 Logical Access audio dataset.

Hsu et al[11] addressed the difficulties in detecting deep fake images created by generative adversarial networks (GANs) and introduced a deep learning approach utilizing pairwise learning. The method employs contrastive loss to differentiate between real and fake images, integrating a reduced DenseNet and a Siamese network structure for feeding pairwise information. Experimental results indicate that this approach surpasses existing fake image detectors. The authors further elaborate on a two-step learning policy for optimizing the Common Fake Feature Network (CFFN) and classifier network independently. They highlight the necessity of a robust backbone network for capturing intricate fake features and discuss the design of the CFFN for detecting both fake faces and general images, achieving enhanced precision and recall. Limitations include the need for retraining with diverse fake features and challenges in gathering those training samples. Future research may focus on extending this methodology to fake video detection.

Zhou et al. [12]introduced a new method for detecting audiovisual deepfakes by assessing both video and audio modalities. It emphasizes the significance of

employing synchronization across both modalities to improve detection accuracy, implying that standard methods that focus primarily on either video or audio are ineffective. The proposed system outperforms independent models and displays high generalization capabilities on previously unseen deepfake types.

Li and LYU et al. [5]proposed a novel deep-learning approach to detect Deep Fake videos by identifying artifacts caused by face warping during the video synthesis process. Unlike traditional methods that rely on a large dataset of both real and fake images for training, this method focuses on the distinctive warping artifacts that result from the limited resolution constraints of Deep Fake algorithms. By employing convolutional neural networks (CNNs), the authors simulate negative examples through simple image processing techniques. Significantly reducing time and resource requirements. The method demonstrates robustness across various datasets, achieving high accuracy in distinguishing real from fake videos, making it a promising tool for combating misinformation and enhancing digital media forensics.

Patel et al.[7] presented an upgraded Dense Convolutional Neural Network (D-CNN) specifically developed for detecting deepfake images, These images are created using generative adversarial networks (GANs). D-CNN architecture is trained on a broad dataset that includes both real and deepfake images, resulting in an accuracy of 97.2%. To improve performance and prevent overfitting, the model's structure contains convolutional, pooling, and fully connected layers, as well as batch normalization and dropout layers. Data augmentation was used to improve generalizability across datasets. The model's performance was examined using measures such as precision, recall, and F1 score, The proposed method shows promising results when compared to previous models. The authors suggested to use of variable input neural networks and effective upscaling techniques to improve performance.

Zhao et al. [12] developed an interpretable spatial-temporal video transformer that decouples spatial and temporal features, enhancing detection robustness while offering insights into model decision-making.

Muppalla et al.[13] proposed a multimodal framework integrating audio-visual features to improve detection accuracy across domains. The proposed method shows enhanced performance in cross-domain testing.

Agarwal et al. [14] presented a biometric-based method for detecting deep-fake videos, specifically face swaps, that combines facial recognition with behavioral analysis of facial expressions and head movements. Their approach uses a convolutional neural network (CNN) to learn behavioral biometrics. Their method demonstrates excellent accuracy and robustness across several datasets while addressing concerns about the exploitation of deepfakes in society.

D. Wodajo et al.[15] propose a Convolutional Vision Transformer (CViT) model for detecting deepfake videos, integrating Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The model achieves 91.5% accuracy on the DeepFake Detection Challenge Dataset.

Dataset used in the study:

To counter this emerging threat, researchers have used different datasets. Some of the most commonly used datasets for deepfake detection using deep learning include DF-TIMIT, UADFV, FaceForensics++, DFDC, Celeb- DF, Google DFD, and WLDR.

Table1. Dataset of images

Dataset	Real Images	Fake Images	Link
DFFD	58,703	240,336	https://github.com/NVlabs/ffhq-dataset
FFHQ	-	70,000 (GAN-based)	https://github.com/NVlabs/ffhq-dataset
iFakeFaceDB	-	87,000 (StyleGAN)	https://github.com/socialabubi/iFakeFaceDB
100k Faces	-	100,000 (StyleGAN)	https://github.com/ozgrozer/100k-faces
ForgeryNet	1,438,201	1,457,861	https://github.com/yinanhe/forgerynet

Table 2. Dataset of Videos

Dataset	Real Videos	Fake Videos	Link
UADFV	49	49	
Deepfake-TIMIT	320	640	https://www.idiap.ch/en/dataset/deepfaketimit
DFFD	1000	3000	https://github.com/NVlabs/ffhq-dataset
Celeb-DF (v2)	590	5639	https://github.com/danmohaha/celeb-deepfakeforensics
DFDC	23,564	104,500	The Deepfake Detection Challenge (DFDC) Preview Dataset
DeeperForensics-1.0	50,000	10,000	https://github.com/EndlessSora/DeeperForensics-1.0
FaceForensic++	1000	4000	https://github.com/ondyari/FaceForensics
ForgeryNet	99,630	121,617	https://github.com/yinanhe/forgerynet
VGGFace2			https://www.tensorflow.org/datasets/catalog/vgg_face2

5. Statement of the Research Problem:

Develop a robust and accurate deep learning model to identify deepfakes to effectively distinguish between authentic and artificially generated facial expressions and movements.

6. Hypothesis:

By leveraging advanced convolutional neural network architectures and multimodal feature extraction techniques, a deep learning model can achieve significantly higher accuracy in detecting deepfake media than existing methods.

7. Research Methodology:

The research will combine theoretical work with hands-on development, involving the following methods:

- **Data Collection and Preprocessing:** Gathering large-scale datasets of real and synthetic media, including images and videos using publicly available sources for research.

- **Model Development:** Implementing deep learning architectures using frameworks TensorFlow / Keras / Pytorch. The model architectures will be evaluated for their ability to detect inconsistencies in deepfake content.

- **Training and Optimization:** Train the model on the dataset and explore the capabilities of different optimization techniques like transfer learning, knowledge distillation, and multi-task learning to improve detection performance.

- **Evaluation:** Testing models on multiple datasets, including both synthetic and real-world deepfake instances, to measure generalization and robustness.

- **Expected Contributions:** A detailed analysis of the state-of-the-art in deepfake detection using deep learning, with a focus on its practical applicability.

8. Significance of the study:

Deep fake detection research employing deep learning is critical because the technique has far-reaching implications in domains such as security, media integrity, personal privacy, and societal trust. This study will helps in : .

1. The preservation of truth and authenticity.

Deepfakes can distort reality and disseminate misinformation and deception. Detecting deep fakes helps to sustain media authenticity, preserving trust in digital content.

2. Mitigation of security threats

Deep fakes are a significant danger to cybersecurity, including impersonation for

fraud, social engineering assaults, and illegal access Detection systems shield individuals and organizations from such threats by detecting altered content.

3. Protection from personal and political harm

Deep fakes are increasingly used for malicious purposes, such as creating fake image or videos to harm individuals' reputations. Accurate detection methods serve to safeguard persons, political stability, and democratic processes.

9. Data Collection Method:

The research require a vast collection of authentic and fraudulent videos. We will collect this dataset from the following places:

1) The Deepfake Detection Challenge Dataset

2) Celeb Deepfake forensics master dataset

[https://github.com/yuezunli/celeb-deepfakeforensics]

3) Kaggle's Deepfake dataset

[https://www.kaggle.com/datasets/vijaydevane/deepfake-detection-challenge-dataset-face-images]

10. Data Analysis:

In the data collection and analysis stage, a dataset with both real and deepfake media will be gathered, and processed beforehand by extracting facial features from each frame.

An important milestone in the study will be understanding the peculiarities of the fake media, be it differences across the modalities like the synchronization between the video and the audio, or spatial identity inconsistencies like the between inner face region and the outer face region. These differences are caused due to the morphing of the original media. The analysis will include the differentiation between the real and the fake media primarily but not limited to the feature extraction, be it facial features, modal, or spatial features. We will leverage the existing feature extraction techniques and explore their ability to identify the alterations to the media. We will evaluate their performance and also experiment with the combination of multiple techniques in order to possibly extract superior performance. We will further feed these extracted features to a CNN-based deep learning architecture, supported by optimization techniques like transfer learning, knowledge distillation, and multi-task learning. We will also explore

the possibility of using architectures like LSTMs and Transformers that are known to perform better on sequential data.

11. Tentative Chater Plan:

Academic research work is presented in several chapters and in this pattern, The following chapter scheme will be adopted by the researcher.

Chapter 1: Introduction:

This chapter introduces research topic, aims and rationale for the study. It describes the methods used. Describe the scope of the study and the specific contributions of research.

Chapter 2: Theoretical Background and Literature Review

This chapter summarizes the existing studies. Review different approaches of deep fake detection. Discuss commonly used datasets and highlight challenges in the field that motivate the proposed work. It also describes deepfake, deep learning and how it is used to detect deepfake.

Chapter 3: Problem Statement and Objectives

This chapter describes the problem statement and objectives of the proposed work.

Chapter 4: Research Methodology:

It explains the methodology used in proposed research.

Chapter 5: Data Interpretation and Analysis:

This chapter present carefully examine the model's predictions on a test dataset to analyse the results of the study, It describes the performance metrics, accuracy, precision, recall, and F1 score.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + FP + TN + FN}}$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}}$$

$$\boldsymbol{F1-score} = \frac{\boldsymbol{2*Precision*Recall}}{\boldsymbol{Precision+Recall}}$$

It compares the results of the proposed method with existing state-of-the-art techniques.

Chapter 6: Findings, Suggestions and conclusion:

This chapter summarizes the main findings, conclusions of the research and suggestions.

12. References and Bibliography:

[1] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," 2018, *arXiv*. doi: 10.48550/ARXIV.1811.00656.

[2] D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Auckland, New Zealand: IEEE, Nov. 2018, pp. 1–6. doi: 10.1109/AVSS.2018.8639163.

[3] S. St *et al.*, "Deep learning model for deep fake face recognition and detection," *PeerJ Comput. Sci.*, vol. 8, p. e881, Feb. 2022, doi: 10.7717/peerj-cs.881.

[4] S. Lyu, "Deepfake Detection: Current Challenges and Next Steps," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, London, UK: IEEE, Jul. 2020, pp. 1–6. doi: 10.1109/ICMEW46912.2020.9105991.

[5] Y. Patel *et al.*, "An Improved Dense CNN Architecture for Deepfake Image Detection," *IEEE Access*, vol. 11, pp. 22081–22095, 2023, doi: 10.1109/ACCESS.2023.3251417.

[6] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

[7] H. F. Shahzad, F. Rustam, E. S. Flores, J. Luís Vidal Mazón, I. De La Torre Diez, and I. Ashraf, "A Review of Image Processing Techniques for Deepfakes," *Sensors*, vol. 22, no. 12, p. 4556, Jun. 2022, doi: 10.3390/s22124556.

[8] L. Guarnera, O. Giudice, and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," in *2020 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition Workshops (CVPRW), Seattle, WA, USA: IEEE, Jun. 2020, pp. 2841–2850. doi: 10.1109/CVPRW50498.2020.00341.

[9] D. Pan, L. Sun, R. Wang, X. Zhang, and R. O. Sinnott, "Deepfake Detection through Deep Learning," in *2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, Leicester, UK: IEEE, Dec. 2020, pp. 134–143. doi: 10.1109/BDCAT50828.2020.00001.

[10] A. Chintha *et al.*, "Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 5, pp. 1024–1037, Aug. 2020, doi: 10.1109/JSTSP.2020.2999185.

[11] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep Fake Image Detection Based on Pairwise Learning," *Appl. Sci.*, vol. 10, no. 1, p. 370, Jan. 2020, doi: 10.3390/app10010370.

[12] Y. Zhou and S.-N. Lim, "Joint Audio-Visual Deepfake Detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, Oct. 2021, pp. 14780–14789. doi: 10.1109/ICCV48922.2021.01453.

[13] S. Muppalla, S. Jia, and S. Lyu, "Integrating Audio-Visual Features For Multimodal Deepfake Detection," in *2023 IEEE MIT Undergraduate Research Technology Conference (URTC)*, Cambridge, MA, USA: IEEE, Oct. 2023, pp. 1–5. doi: 10.1109/URTC60662.2023.10534969.

[14] H. Agarwal, A. Singh, and R. D, "Deepfake Detection Using SVM," in *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India: IEEE, Aug. 2021, pp. 1245–1249. doi: 10.1109/ICESC51422.2021.9532627.

[15] D. Wodajo and S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer," 2021, *arXiv*. doi: 10.48550/ARXIV.2102.11126.

[16] https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

[17] https://www.techtarget.com/whatis/definition/deepfake

[18] https://dl.acm.org/doi/fullHtml/10.1145/3491102.3517446

[29] https://seon.io/resources/dictionary/deepfakes/

[20] https://vlinkinfo.com/blog/top-ai-deepfake-detector-tools/

[21] https://www.britannica.com/technology/deepfake

Working plan:

Year 1:

- **First 6 month's plan:**
    - Collection of previous studies and books on research topic and write review and publication.

- **Second 6 month's plan:**
    - Study of deep learning framework.

Year 2:

- **First 6 month's plan:**
    - Collection of appropriate datasets and start to design the system.

- **Second 6 month's plan:**
    - Implementation the proposed system.

Year 3:

- **First 6 month's plan:**
    - Start the data analysis and write the chapter Data Collection and Analysis and publication

- **Second 6 month's plan:**
    - Publication of paper and Report writing, printing, binding and submission of thesis

Research Student **Research Supervisor**

Mrs. Sudha Pradeep Patil **Prof. Dr. Sudhir B. Jagtap**